マルコフモデルによるかな漢字文候補の絞り込み

荒木 哲郎
福井大学 工学部

池原 悟
NTT コミュニケーション科学研究所

真田 陽一
福井大学 工学部

間瀬 恒志
福井大学 工学部

1 はじめに

日本語文は、通常約3000種の文字(特に漢字文字)を用いて書かれるため、コンピュータのファイルに入力することが容易ではない。べた書きのかな文を漢字かな混じり文に変換する方法については、これまでにもいろいろ研究されているが、同音語による曖昧さと分かち書き処理の曖昧さを同時に解決しなければならず、現在のところではまだ十分な精度を得るには至っていない。

これまでに日本語の統計的な情報を用いた日本語処理の研究の一つとして、日本語の文字間の結合性に着目した方法、すなわち文節単位のべた書きかな文字列から変換された大量の漢字かな混じり候補を、文節内の漢字かな文字のマルコフモデル(文節漢字かなマルコフモデル)を用いて絞り込む方法が提案され、その有効性が示されている[2][3]。また文節マルコフモデルを用いて絞り込まれた漢字かな混じり文節候補を、相互に組み合わせて作られる文候補ラティスから、最も確からしい文候補を文漢字かなマルコフモデルを用いて絞り込む方法の有効性が報告されている[4]。

本論文では、漢字かなマルコフモデルに加えて、さらに文を構成する単語列に対する品詞列のマルコフモデル(品詞マルコフモデル)を考え、これらを組み合わせた絞り込み方法を提案する。

2 基本的な定義と2重マルコフモデルを用いた文候補絞り込み方法

2.1 基本的な定義

文字列 $\Gamma = s_1 s_2 \cdots s_l$ によって表現される日本語文は、そのすべての要素 (s_i) がかな文字であるとき、かな文と呼ばれる。また文字列 のすべての要素 (s_i) が漢字またはかな文字であるとき、その日本語文を漢字かな混じり文と呼ぶ。かな文 Γ に対する日本語の漢字かな混じり文を、$\Lambda(\Gamma) = t_1 t_2 \cdots t_m$ によって表現する。

日本語の文は、文節と呼ばれる構文の単位に分割できるとする。かな文節から、漢字かな混じり文節候補の生成方法は、[4] に従う。

日本語のかな文 Γ が文節 $\gamma_1, \gamma_2, \cdots,$ および γ_n に分割されるとき、対応する漢字かな混じり文 $\Lambda(\Gamma)$ はそれぞれ文節 $\lambda(\gamma_1), \lambda(\gamma_2), \cdots,$ および $\lambda(\gamma_n)$ に分割され、また$\Lambda(\Gamma)$ を構成する単語列に対する品詞情報(品詞コード)列を、$\Theta(\Lambda(\Gamma)) = u_1 u_2 \cdots u_n$ によって表現する。。

最初に、漢字かな混じり文節候補、または文候補の最適性を評価するのに用いられるマルコフモデルを定義する。

【定義 1】 各漢字かな混じり文節が、$\lambda(\gamma_k) = t_{v(1)} t_{v(2)} \cdots t_{v(r)}$ によって与えられ、その$\lambda(\gamma_k)$ の品詞列が$\theta(\lambda(\gamma)) = u_{w(1)} u_{w(2)} \cdots u_{w(s)}$ によって表されるとき、 全ての文節に対する 2重漢字かなマルコフ連鎖確率 $p(t_j|t_{j-2}t_{j-1})$ ($v(1) \leq j \leq v(r)$) および $p(u_i|u_{i-2}u_{i-1})$ ($w(1) \leq i \leq w(s)$) の集合をそれぞれ KBM および HBMと呼ぶ。. 同様に各漢字かな混じり文が、$\Lambda(\Gamma_k) = t_1 t_2 \cdots t_m$と与えられ、その$\Lambda(\Gamma_k)$ 品詞列が$\Theta(\Lambda(\Gamma_k)) = u_1 u_2 \cdots u_n$によって表されるとき、全ての文に対する2重漢字かなマルコフ連鎖確率 $p(t_j|t_{j-2}t_{j-1})$ ($1 \leq j \leq m$) および $p(u_i|u_{i-2}u_{i-1})$ ($1 \leq i \leq n$) の集合をそれぞれ KSMおよび HSMと呼ぶ。ここで、2つの空白記号 (␣と表す) が、各文および各文節の先頭、末尾に付加される。

【定義 2】漢字かな混じり文節候補および漢字な混じり文候補 を、$\lambda(\gamma_k) = t_{v1} t_{v2} \cdots t_{vr}$ および$\Lambda(\Gamma) = t_1 t_2 \cdots t_m$ 、また$\lambda(\gamma_k)$ および $\Lambda(\Gamma))$ を構成する単語列に対応する品詞列をそれぞれ、$\theta(\lambda(\gamma)) = u_{w(1)} u_{w(2)} \cdots u_{w(s)}$および $\Theta(\Lambda(\Gamma_k)) = u_1 u_2 \cdots u_n$と表すとき、文節候補または文候補に対して、$KBM$, HBM, KSM

および HSM を用いて、計算される次式の値 Cを、それぞれ文節漢字かなコスト (K_B)、文節品詞コスト (H_B)、文漢字かなコスト (K_S) および文品詞コスト (H_S) と呼ぶ。

$$C = -\sum_{i=1}^{n+2} \log_2 p(x_i|x_{i-2}x_{i-1})$$

ここで x_j は、$j<0$ または $j>n$ のとき、空白記号 $\sqcup$ を表す。

各文節γ_kに対する可能な全ての漢字かな混じり文節候補の集合を $\Omega(\lambda(\gamma_k))$ と表すとき、文節漢字かなコスト K_B 値を用いて、$\Omega(\lambda(\gamma_k))$ の中の最適な漢字かな混じり文節候補を絞り込む方法はすでに提案されている。その結果によれば、正しい漢字かな混じり文節が第一位候補に含まれる割合は、83.7(標本外データ) − 98.2 %(標本内データ) である。

この結果を文候補の生成に適用するために、第一位から第十位までの文節候補の集合を用いて構成される文候補ラテイスを、次のように定義する。ここで文節コスト C の値を用いて、漢字かな文節 $\Omega(\lambda(\gamma_k))$ から絞り込まれた第一位から第十位までの候補の集合を、$\Omega(\lambda(\gamma_k))^{(10)}$ と表す。

【定義 3】[1]日本語のかな文 Γ が文節 γ_1, γ_2, $\cdots$, および γ_nに分割され (すなわち $\Gamma = \gamma_1\gamma_2\cdots\gamma_n$)、かな文に対する漢字かな文 $\Lambda(\Gamma)$ が、文節$\lambda(\gamma_1)$, $\lambda(\gamma_2)$, $\cdots$, および $\lambda(\gamma_n)$、に分割されているとき、第一位から第十位までの漢字かな文節候補$\Omega(\lambda(\gamma_k))^{(10)}$ $(1 \leq k \leq n)$ の列を、漢字かな混じり文候補ラテイスとよび、次のように表す。$L(\Lambda(\Gamma)) = \Omega(\lambda(\gamma_1))^{(10)}\Omega(\lambda(\gamma_2))^{(10)}\cdots\Omega(\lambda(\gamma_n))^{(10)}$.

但し、正しい漢字かな混じり文節は、常に文節候補の各集合$\Omega(\lambda(\gamma_k))^{(10)}$ に含まれているものと仮定する。

文候補ラテイス $L(\Lambda(\Gamma))$ の例を図 1 に示す。

2.2 品詞情報の 2 重マルコフモデルを用いた文候補絞り込み方法

本章では、文候補ラテイスから得られる漢字かな混じり候補の内、最適な文を選択する方法を定義する。

【1. 品詞コストのみを用いた文候補絞り込み法】

文品詞コスト H_Sが最小の値を持つ文を最適な文候補として、絞り込む方法を H_S-方法と呼ぶ。また文の中の文字数 L によって正規化された最小の文品詞コスト H_Sの値 (すなわち、H_S/L) を持つ文をであると定義する方法を、NH_S-方法と呼ぶ。さらに文品詞コスト H_Bおよび文節品詞コスト H_Sを加算したコスト値を用いて絞り込む方法を、$(H_S + H_B)$-方法と呼ぶ。

【2. 品詞コストと漢字かなコストを併用した文候補絞り込み法】

文候補ラテイスから得られる最適な漢字かな混じり文候補を、文漢字かなコスト K_Sと文節漢字かなコスト K_Bおよび文品詞コスト H_S、文節品詞コスト H_Bの和が最小な値を持つ文をであると定義する方法を、$(K_S+K_B+H_S+H_B)$-方法と呼ぶ。また文および文節の漢字かなコスト K_Bおよび K_Sの加算で絞り込みを行なう方法を $(K_S + K_B)$-方法、また文および文節の品詞コスト $H_S + H_B$の加算で絞り込みを行なう方法を $(H_B + H_S)$-方法と呼ぶ。

3 実験結果

3.1 実験条件

1. マルコフ連鎖確率の統計に用いられる総文節数: 70 日分の新聞記事データで、283,963 文節数
2. 一文節当たりの平均文字数:6
3. 三つの文候補絞り込み方法を、評価するのに用いられる文、漢字かな文字、文節の数:
   (a) 文の数: 200 文 (標本外データ), 214 文 (標本内データ)
   (b) 文節総数: 2041 文節 (標本外データ), 1899 文節 (標本内データ)
   (c) 平均の文の長さ: 31.3 文字 (標本外データ), 27.8 文字 (標本内データ)
4. 辞書の単語数:430,000 語

3.2 実験結果

2 章で述べた文節および文の品詞のコストのみによる文候補絞り込み法を用いて、選択された第一位から第十位までの候補の中に 、正しい漢字かな混じり文が含まれる割合 (累積正解率) を求めた実験結果を、図 2-(a)(標本内データ) および図 2-(b)(標本内データ) に示す。ま

[1]この実験においては、日本語音声出力システムから得られる読みの情報 [4] が、文節コストの他に文節候補の絞り込みに用いられている。これらの情報を用いた文節候補の絞り込みを行なうことによって、正解率が改善され、多くの正しい漢字かな混じり文節候補が、10 位より小さい順位の候補に含まれれることに注意する。

た文節および文の漢字かなコスト、および品詞コストを加算した文候補絞り込み法を用いて得られた実験結果を、図 2-(a)(標本内データ) および図 2-(b)(標本内データ) に示し、また第一位から第十位までの候補の例を図 4 に示す。

[1] 品詞コストのみを用いた文候補絞り込み法による累積正解率の比較:

1. (H_S+H_B)-方法は、H_S-方法よりも第一位正解率で 6.1(標本内)-10.5(標本外) %、また 10 位内累積正解率で 2.8 %(標本内)-13.5 %(標本外) 優れている。このことは文候補ラテイスから、漢字かな混じり文候補を正しく選択する際に、文コストと文節コストの両方の値の組合せて評価する方が、文コスト単独で選択するよりも有効であることを意味している。

2. NH_S-方法が H_S-方法よりも、10 位内累積正解率で 5%(標本外)-12.6 %(標本内) 悪い。これはいろいろな長さ (文を構成する単語数に対応する品詞の数) の文候補が、文品詞コストの値を用いて評価されるとき、文品詞コストは文内の品詞数によって正規化されない方が有効であることを意味している。

[2] 漢字かなコストと品詞コストの加算による文候補絞り込み法の比較:

図 3 より、漢字かなコストと品詞コスト加算した絞り込み方法は、従来の漢字かなコストを用いた絞り込み方法に比べて、第一位正解率で 0.9(標本内) − 8.8(標本外)%優れていることがわかる。

4 おわりに

本論文は、文節マルコフモデルによって決定された漢字かな混じり文節候補の組合せから構成される漢字かな混じり文候補ラテイスより、文マルコフモデルを用いて文候補を正しく選択する方法を提案した 。

実験結果から、以下の結論を得た。

1. (H_S+H_B)-方法は、S-方法よりも 6.1-10.5 %優れている。

2. NH_S-方法は H_S-方法よりも 悪い。

3. 漢字かなおよび品詞コストの両方の値に基づいた文候補絞り込み法は、従来の漢字かなコストを用いた絞り込み方法に比べて、第一位正解率で 0.9(標本内) − 8.8(標本外)% 向上することがわかった。

これらの結果から、2 重マルコフモデルを用いた文候補絞り込み法は、曖昧な漢字かな混じり文節候補から構成される文候補ラテイスより漢字かな混じり文候補を正しく選択するのに有効であることがわかった。

今後は音声入力などの曖昧な音節認識候補ラテイスから、漢字かな混じり文を正しく絞り込む方法について、さらに研究することなどがあげられる。

参考文献

1. 宮崎、大山: "日本文音声出力のための言語処理方式", 情報処理, Vol.27, 11, pp1053-1061 (1986)

2. 荒木、池原、芳永、真田: " マルコフ連鎖モデルによる文節かな漢字変換候補の絞り込み方法", 情処ＮＬ研究会, Vol.99-6, pp41-48 (1994)

3. 荒木、池原、横川、真田: " 日本語文音声出力からの読み情報を用いた漢字かな混じり文節候補の絞り込み", 情処ＮＬ研究会, Vol.102-15, pp113-120 (1994)

4. 荒木、池原、真田、芳永: " マルコフモデルを用いて漢字かな混じり文候補を選択する方法", 情処ＮＬ研究会, Vol.105-2, pp7-13 (1995)

入力文 さこくの／ゆめである

候補順位		
1	鎖国の (1230 7410)	夢である (1100 7255 2780 2786)
2	さ刻の (1100 1100 7410)	輸めである (2430 1100 7255 2780 2786)
3	さ告の (1100 1920 7410)	結めである (2390 1100 7255 2780 2786)
		努である (4100 7055 2780 2786)
		ゆめである (1100 7255 2780 2786)
（ＸＸＸＸ）品詞コード		ユメである (1100 7255 2780 2786)

図１　文節マトリックスの例

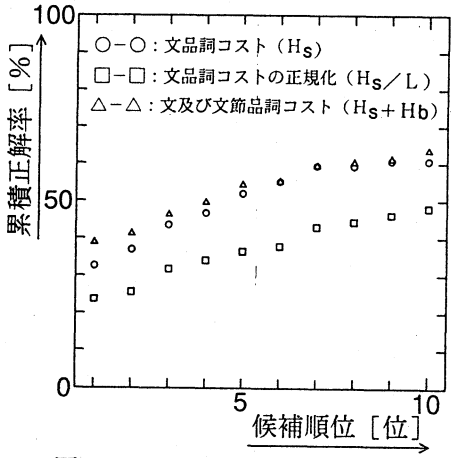

図2−a　品詞コストのみによる絞り込み（標本内）

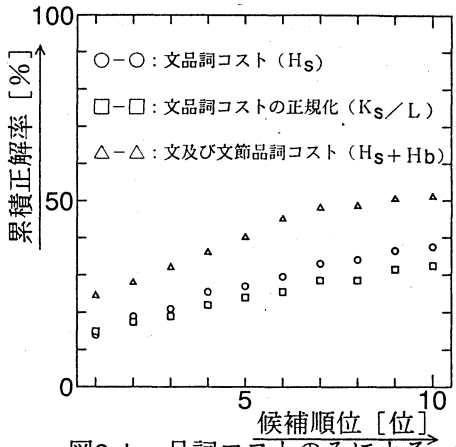

図2−b　品詞コストのみによる絞り込み（標本外）

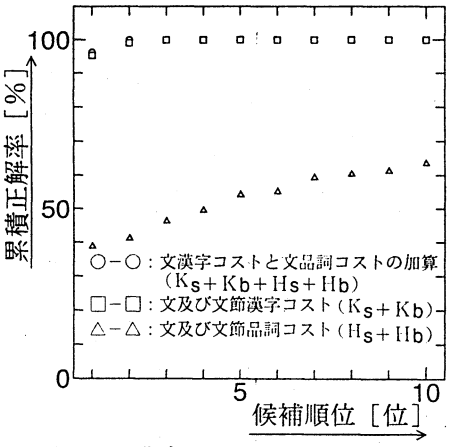

図3−a　漢字コストと品詞コストの加算による絞り込み（標本内）

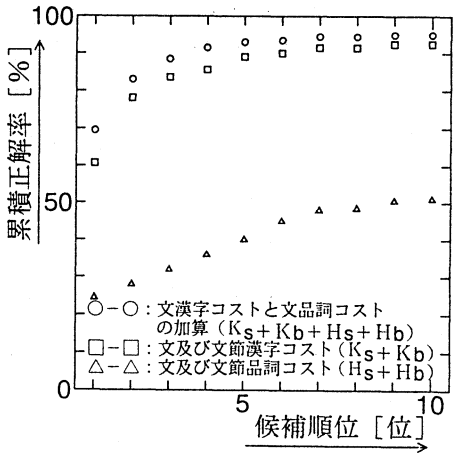

図3−b　漢字コストと品詞コストの加算による絞り込み（標本外）

入力文：さこくのゆめである 正解文候補：鎖国の夢である		品詞コード 1230 7410 1100 7255 2780 2786	品詞コスト 15.451542
候補順位	1	さ刻の夢である	1100 1100 7410 1100 7255 2780 2786	14.647322
	2	さ刻のゆめである	1100 1100 7410 1100 7255 2780 2786	14.647322
	3	さ刻のユメである	1100 1100 7410 1100 7255 2780 2786	14.647322
	4	鎖国の夢である	1230 7410 1100 7255 2780 2786 名詞 各助詞 名詞 助動詞 動詞 動詞	15.451542
	5	鎖国のユメである	1230 7410 1100 7255 2780 2786	15.451542
	6	鎖国のゆめである	1230 7410 1100 7255 2780 2786	15.451542
	7	さ告の夢である	1100 1920 7410 1100 7255 2780 2786	25.308439
	8	さ告のゆめである	1100 1920 7410 1100 7255 2780 2786	25.308439
	9	さ告のユメである	1100 1920 7410 1100 7255 2780 2786	25.308439
	10	さ刻の努である	1100 1100 7410 4100 7255 2780 2786	1012.208277

図4　入力文に対する文候補の具体的な例